Application of NGS-generated SNP data to complex crops studies: the example of *Musa* spp. (banana)

Hueber Y[1], Sardos J[1], Hřibová E[2], Van den houwe I[3], Roux N[1], Rouard M[1]

[1] Bioversity International - Parc Scientifique Agropolis II, 34397 Montpellier, France
[2] Institute of Experimental Botany - Rozvojová 263, 165 02 Prague, Czech Republic
[3] Bioversity International Transit Center - Willem de Croylaan 42 box 2455, BE3001 Heverlee BE3001, Belgium

Abstract: Over recent years, Next-Generation Sequencing (NGS) has become widely used by scientists to obtain high resolution genomic data from many species. In parallel, different Next-Generation genotyping techniques have been developed. However, determining their respective suitability to different research contexts can be challenging, especially when studying highly heterozygous and/or allopolyploid plants. Genotyping data was generated from diploids and triploids *Musa* (banana) samples using two restriction-site associated DNA sequencing methods: Genotyping by Sequencing (GBS) and Restriction site Associated DNA markers (RAD). Both methods reduce the complexity of the targeted genomes and allow obtaining millions of markers across many individuals at a reasonable cost. The potential and limitations of both techniques have been highlighted based on different concrete applications such as: i) Filtering of raw data to get highly reliable markers for Genome Wide Association Studies (GWAS), ii) Comparison of the number, depth and distribution of markers obtained from each technique using 12 diploid and 4 triploid accessions, iii) Computation of phylogenetic trees for genetic diversity analyses involving taxonomically distant taxa, and iv) Use of low quality DNA from leaf samples gathered during collecting missions.

8 steps to get highly reliable markers for GWAS

 # Raw variants (SNPs, short indels)

148,108

1) Remove individuals with missing data > 50 %
2) Discard markers with one or more missing genotypes

46,418

3) Remove non-polymorphic markers
4) Keep only biallelic markers

22,456

5) Remove markers with Fis (inbreeding coefficient) score outside normal range of gaussian distribution (in our case inferior to -0,8)

21,769

6) Keep markers with minor allele frequency (MAF) ≥ 5 %
7) Set to missing genotypes positions with read depth < 10
8) Discard markers > 9 missing genotypes

5,544

 # Analysis-ready variants

Figure 1: Filtration pipeline on raw variants (SNPs, short indels) called on 106 accessions of *Musa* from the International Transit Center (ITC) using Genotyping-By-Sequencing (GBS) single-end methodology to get highly reliable markers for Genome Wide Association Studies (GWAS). Raw variants were called before filtration using the GBS analysis pipeline (TASSEL Version 3). Starting with 148,108 raw variants, only 5,544 biallelic variants were used to perform GWAS analysis.

GBS vs RAD

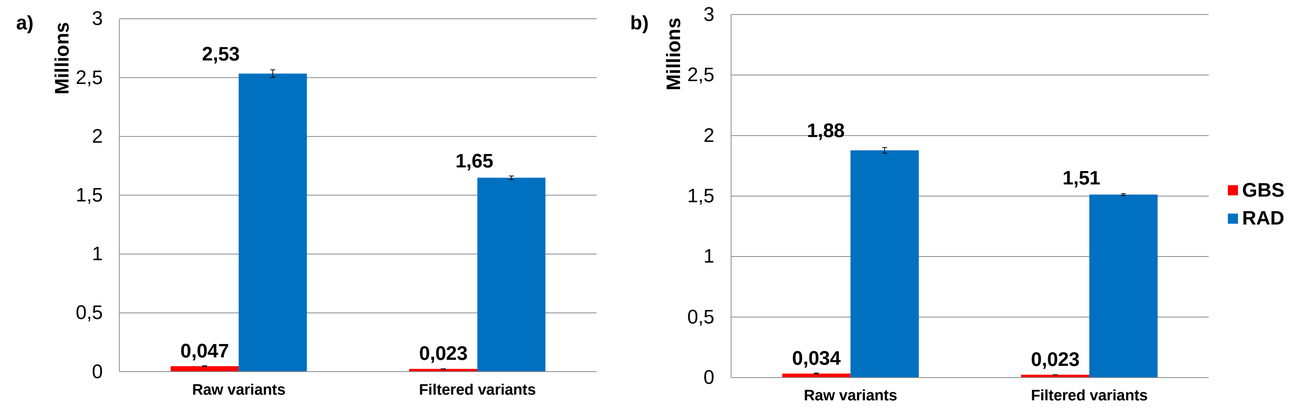

Figure 2: Raw and filtered marker (SNPs, small indels) average number for a) 12 diploids and b) 4 triploids from ITC. Filtering process includes steps 2, 3, 5, 7 and 9 (adapted to the number of individuals: maximum 4 and 1 missing genotypes for diploids and triploids, respectively) from Figure 1. Error bars represent standard error.

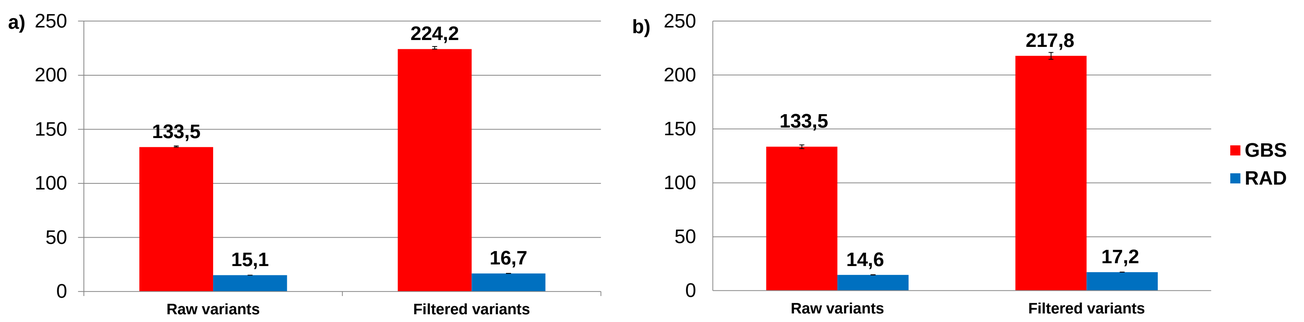

Figure 3: Depth of coverage average of raw and filtered markers for a) 12 diploids and b) 4 triploids from ITC. Error bars represent standard error.

Use of low quality DNA extracted from leaf samples

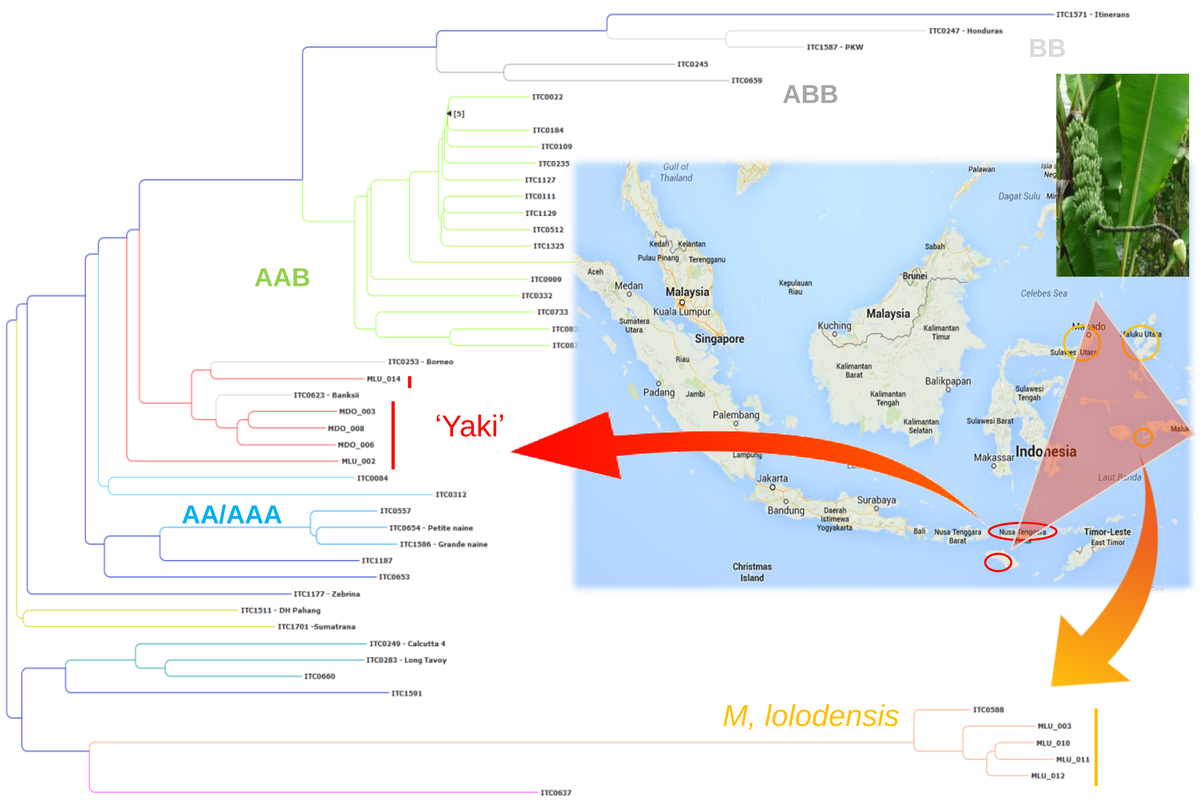

Figure 6: SNP genotypes obtained from RAD performed on low quality DNA were successfully used to locate wild samples obtained from a collecting mission within a diversity tree. Leaf samples were collected in remote areas of the Indonesia "triangle", conserved in a cooling box before being sent by DHL at room temperature to Czech Republic where DNA was extracted at the Institute of Experimental Botany (IEB). The obtained DNA, partly to completely degraded, did not pass quality control standard for RAD sequencing. However, the data obtained were suitable for a phylogenetic analysis.

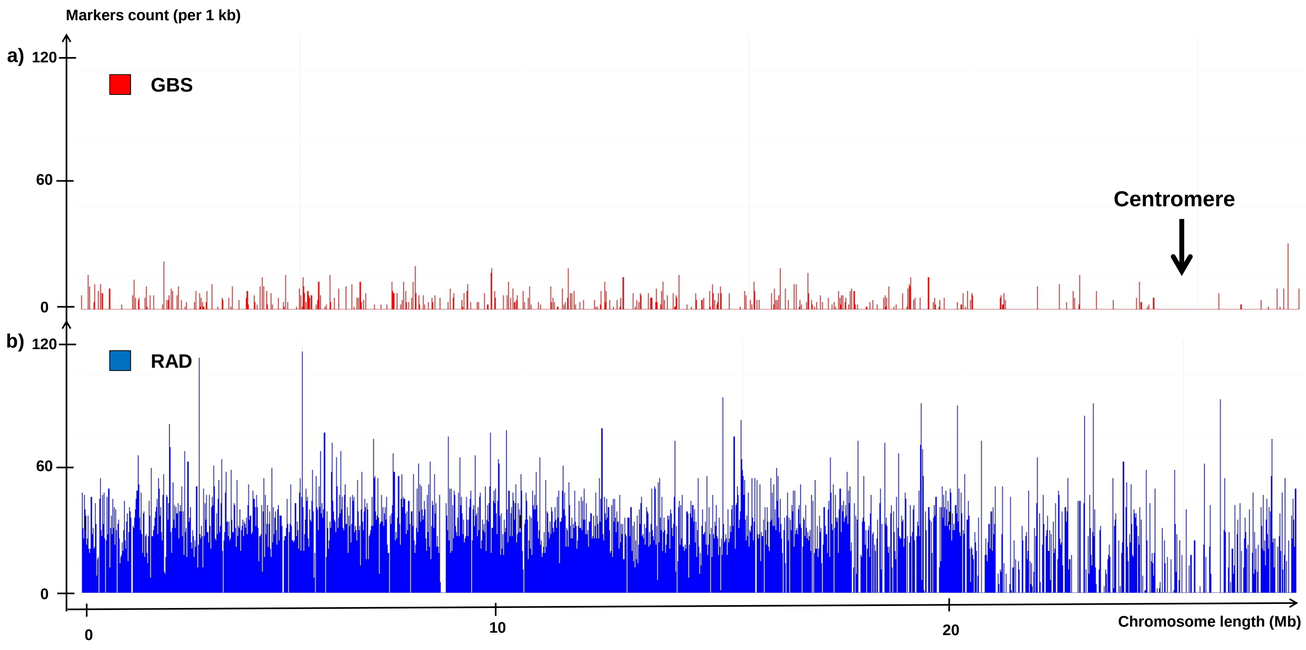

Figure 4: Markers count (filtered markers from the 12 diploids) per interval of 1 kb along chromosome 1 with a) GBS (red) and b) RAD (blue) methodology.

Note: 48-plex GBS was performed at Cornell University (USA) using a protocol modified from *Elshire et al., 2011*. RAD sequencing (paired end reads ~1,5x) was performed at the Beijing Genomics Institute (China).

Genetic studies

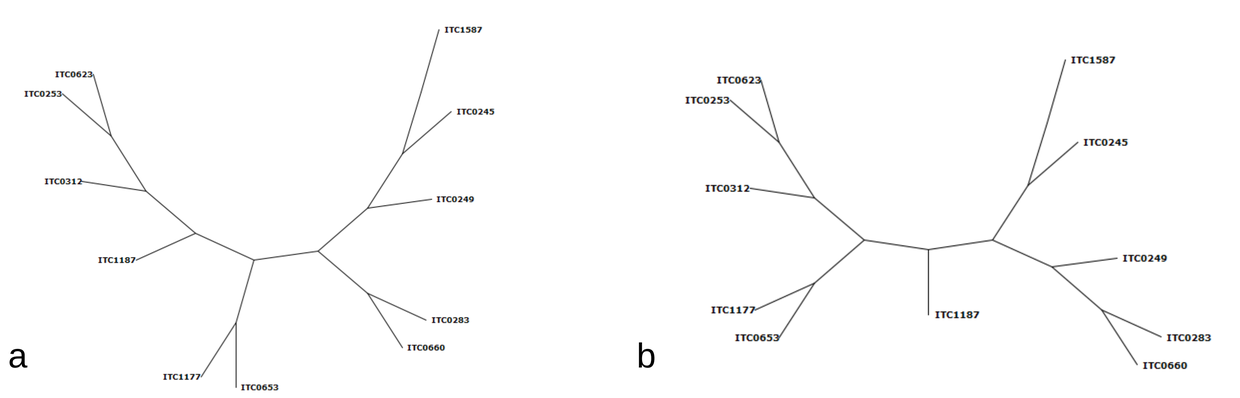

Figure 5: Phylogenetic trees generated with markers coming from a) GBS (3257 SNPs) and b) RAD sequencing (12880 SNPs) on 11 Musa diploids. SNP were filtered with missing value <50%, MAF ≥ 0.05, coverage >20, LD < 0.5 . Trees were computed as described in SNPhylo (Lee et al, 2014) but using PhyML (best of NNI/SPR). Overall, the results are similar but topology differs for some branches highlighting the importance of sampling for subspecies resolution despite a high number of data points.

References:

Elshire LG, Glaubitz JC, Sun Q, Poland JA, Kawamoto K et al.(2011) A Robust, Simple Genotyping-by-Sequecing (GBS) Approach for High Diversity Species. PLoS ONE 6(5):e19379. doi:10,1371/journal.pone.0019379

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol Biol *772,* 157-78

Acknowledgements:

Noel Chen and He Qiongzhi (BGI)

Katie Hyma and Sharon Mitchell (Cornell University)